УДК 004.81

Ю.Б. Крапивин
Брестский государственный технический университет, Беларусь
Беларусь, 224017, г. Брест, ул. Московская, 267

Автоматический поиск заимствованных из Интернет-источников фрагментов

Y.B. Krapivin
Brest State Technical University, Belarus
Belarus, 224017, c. Brest, Moskovskaja st., 267

Automatic Retrieval Adopted from the Internet-Resources Fragments

Ю.Б. Крапивін
Брестський державний технічний університет, Білорусь
Білорусь, 224017, м. Брест, вул. Московська, 267

Автоматичний пошук запозичених з Інтернет-джерел фрагментів

В статье рассмотрено решение задачи автоматического распознавания фрагментов текстового документа, заимствованных из Интернет-доступных источников. Приведена постановка задачи автоматического распознавания плагиата, дается описание системы автоматического распознавания воспроизведенных фрагментов текстовых документов, а также соответствующего алгоритма их обнаружения. Предложенные решения позволяют системе, построенной в соответствии с ними, распознавать случаи заимствования как из заранее заданной полнотекстовой базы данных, так и из полнотекстовой базы данных, полученной в результате автоматического Интернет-поиска по ключевым словам, выделенным из анализируемых документов.
Ключевые слова: естественный язык, автоматическая обработка текстов, заимствованный фрагмент, ключевые слова.

In the article, automatic recognition of the fragments of the text documents adopted from the Internet-available resources is considered. The article presents a definition of plagiarism identification problem, description of the system of the automatic recognition of reproduced fragments of the text documents, as well as the corresponding algorithm of their detection. The submitted solutions allow the system, which is built according to them, to recognize the adoptions both from the given full-text database and full-text database, created due to the automatic Internet search with the key words, marked out from the reviewed documents.
Key Words: natural language, automatic text processing, adopted fragment, key words.

У статті розглянуто рішення задачі автоматичного розпізнавання фрагментів текстового документа, запозичених з Інтернет-доступних джерел. Наведена постановка завдання автоматичного розпізнавання плагіату, дається опис системи автоматичного розпізнавання відтворених фрагментів текстових документів, а також відповідного алгоритму їх виявлення. Запропоновані рішення дозволяють системі, побудованій відповідно до них, розпізнавати випадки запозичення як із заздалегідь заданої повнотекстової бази даних, так і з повнотекстової бази даних, отриманої в результаті автоматичного Інтернет-пошуку за ключовими словами, виділеними з аналізованих документів.
Ключові слова: природнича мова, автоматична обробка текстів, запозичений фрагмент, ключові слова.

Введение

Развитие информационных технологий, обеспечивающих электронную форму хранения подавляющего большинства текстовых документов практически во всех об-

ластях человеческой деятельности, существенно обострило проблему их качественной компьютерной обработки с целью автоматизации решения различных прикладных задач. Одной из них является трудоемкая задача определения плагиата в текстовых документах.

Постановка задачи

Существует множество определений плагиата. В русском языке значение слова плагиа́т (от лат. plagio – похищаю) известно давно и с течением времени не претерпело значимых изменений.

Чаще всего под плагиатом понимают умышленное присвоение авторства на чужое произведение литературы, науки, искусства, изобретение или рационализаторское предложение (полностью или частично). Предусматривается уголовная и гражданская ответственность за нарушение авторских и изобретательских прав [1].

Случаи плагиата могут быть и непреднамеренными, например, вследствие сильного внешнего информационного влияния, которое может проявляться в использовании идей или характерного способа их выражения, а также несоблюдения общепринятых правил цитирования, если речь идет об информации, представленной в текстовой форме.

Способы обнаружения плагиата варьируются в зависимости от того, в какой предметной области рассматривается данное понятие. Далее будем исследовать задачу распознавания плагиата применительно к информации, представленной в виде текстовых документов на естественном языке (ЕЯ).

Задачу распознавания плагиата в контексте её приложений и сложности решения мы рассматриваем в двух постановках:

– распознавание заимствованных (воспроизведенных) фрагментов текста (точное совпадение или совпадение с точностью до лексической и грамматической синонимии);

– распознавание семантически эквивалентных фрагментов, по крайней мере, на уровне основных типов знаний о внешнем мире / предметной области, а именно объектов (концептов), фактов (семантических отношений между объектами типа С-А-О, где С – субъект, А – акция, О – объект) и причинно-следственных отношений между самими фактами, полными и неполными, которые отображают закономерности внешнего мира / предметной области [2], [3]. В определённом смысле вторую постановку задачи можно рассматривать как развитие первой, которой и посвящена настоящая работа. Таким образом, речь идет о распознавании воспроизведенных фрагментов текстовых документов, т.е. тех фрагментов данного (входного) документа, которые заимствованы из других документов, представленных, в конечном счете, в некоторой заданной многоязычной полнотекстовой базе данных, в нашем случае – белорусско-русской.

В настоящее время существуют некоторые системы, решающие такого же типа задачи. Наибольшее распространение получили среди них системы WCopyfind, CopyCatch, PlagiatInform, Анти-Плагиат, оперирующие алгоритмами распознавания явного, но не всегда точного заимствования фрагментов текста: их соответствие по лексическому составу и позициям лексических единиц, либо только по лексическому составу, с учётом простейших морфологических преобразований и отношений синонимии. К тому же, каждая из этих систем поддерживает работу только с одним языком. Существующие системы в большинстве своем не обеспечивают приемлемых результатов работы по таким показателям, как полнота и точность анализа текстов, скорость их обработки, объемы используемой памяти ЭВМ, что во многом связано с недостаточной эффективностью реализуемых алгоритмов [4].

Структурно-функциональная схема системы

В работе [4] определена базовая функциональность, а также структурно-функциональная схема системы автоматического распознавания воспроизведенных фрагментов текстового документа, которая, в качестве основных, включает подсистемы: определения языка текстового документа, машинного перевода, автоматического индексирования и поиска релевантных документов, а также распознавания эквивалентности фрагментов документов.

Наличие подсистем определения языка текстового документа и машинного перевода обусловлено тем, что рассматриваемая задача решается в многоязычной информационной среде.

Для определения языка текстового документа применялись методы, ориентированные на использование знаний о естественном языке в пределах от уровня алфавита до лексико-грамматического уровня глубины ЕЯ [5].

В качестве подсистемы МП использовалась уже существующая система машинного перевода в белорусско-русской информационной среде [2], [6]. Это система трансферного типа, кроме того, она «умеет» настраиваться на предметную область на основе автоматического анализа предлагаемого пользователем соответствующего корпуса текстов.

Подсистема автоматического индексирования и поиска релевантных документов обеспечивает возможность поиска документов, релевантных входному, в заранее заданной полнотекстовой БД, и Интернет-поиска по ключевым словам, автоматически выделенным из анализируемого документа.

Что касается функциональности собственно распознавания воспроизведенных фрагментов текстовых документов, то она ориентирована не только на явное, но и неявное заимствование с точностью до парадигм лексических единиц и отношений лексической и грамматической синонимии.

Общая функциональность системы потребовала в совокупности использования развитого лингвистического процессора (ЛП), ориентированного на автоматический лексико-грамматический, синтаксический и семантический уровень анализа и синтеза языка. Такой сложный базовый модуль системы опирается в своей работе на лингвистическую базу знаний (ЛБЗ), включающую различные, в том числе и эталонные, словари языков и корпусы их текстов, грамматики языков, классификаторы их свойств на различных уровнях глубины языков, так называемые распознающие лингвистические модели анализа текста в виде разработанных экспертом лингвистических правил (паттернов) и т.д. [2].

Подсистема автоматического индексирования и поиска релевантных документов

В рамках подсистемы автоматического индексирования и поиска релевантных документов решается задача отбора документов, релевантных входному, для последующего анализа на предмет наличия в нем заимствований из полученного множества. То есть указанная задача включает следующие подзадачи: поиска релевантных документов, создания их полнотекстовой БД и обнаружения заимствованных фрагментов. При этом релевантными считаются документы, возвращаемые информационно-поисковой системой Google [7] в качестве ответа на поисковый запрос в виде ключевых слов, автоматически выделенных из анализируемого документа.

Процесс выделения ключевых слов из входного документа заключается в назначении весовых коэффициентов нормализованным словам – термам, составляющим входной документ, а также в отборе требуемого их количества среди тех, чей вес превышает заданное пороговое значение. Расчет весовых коэффициентов осуществляется по методу TF-IDF [8], учитывающему статистическую информацию о вхождениях слов как в анализируемый документ, так и в корпус текстов.

Таким образом, вес w_{dk} k-го терма входного документа d рассчитывается по формуле:

$$w_{dk} = TF \cdot IDF \ ,$$

где TF – частота терма в анализируемом документе d, $TF = \frac{n_k}{N}$, n_k – число вхождений k-го терма во входной документ, а N – общее число всех термов документа; IDF – обратная частота документа, $IDF = \log \frac{N_{DB}}{N_k}$, N_{DB} – число документов в корпусе, а N_k – число документов корпуса, содержащих k-й терм.

В качестве корпуса текстов для расчета обратной частоты документа используется полнотекстовая база данных эталонных документов. Существует также возможность изменять как количество ключевых слов в запросе, так и количество документов, получаемых при проведении Интернет-поиска. Вполне удовлетворительные результаты работы подсистемы достигаются в случае запроса из 15 ключевых слов с сохранением первых 50 Интернет-доступных документов.

Важно отметить, что задача нормализации слов, т.е. их приведения к каноническому виду, решается путем использования функциональности, предоставляемой подсистемой МП, опирающейся на ЛП и ЛБЗ, включающую многочисленные словари, в том числе и базовый словарь русского языка, содержащий слова, сгруппированные по словоизменительным парадигмам. В базовом словаре парадигма представлена совокупностью словоформ совместно с соответствующими им лексико-грамматическими кодами (ЛГК). ЛГК отражает принадлежность слов лексико-грамматическим классам или, иначе, частям речи, (существительное, прилагательное, глагол и т.д.) и подклассам (например, личные местоимения, возвратные местоимения и т.д.) в соответствии с лексико-грамматическим классификатором, также являющимся компонентом ЛБЗ. Каждая парадигма начинается с канонической формы – словоформы, которая условно считается основной (первой). Например, каноническая форма для имени существительного – именительный падеж единственного числа; для глагола – неопределенная форма глагола. Однако возможны ситуации, когда одна и та же словоформа присутствует в нескольких парадигмах, и в этом случае выбор канонической формы для такой словоформы входного документа является неоднозначным. Поэтому имеет место лексико-грамматический бесконтекстный анализ входного текста [6], позволяющий определить среди возможных вариантов лексико-грамматического анализа предложения наиболее вероятные, т.е. однозначно установить ЛГК, а значит найти соответствующую парадигму и выделить в ней каноническую форму слова.

Алгоритм распознавания заимствованных предложений входного текста

В основу эффективного решения рассматриваемой задачи положен следующий разработанный нами алгоритм распознавания заимствованных из текстовых документов БД отдельных предложений:

1. Начало.

2. Построение обратного индекса I_T входного текста T: выбор из T множества всех попарно различных канонических слов, т.е. построение словаря W_T канонических слов входного текста, с указанием для каждого слова $w_i \in W_T$ множества N_i всех номеров тех предложений из T, в которых это слово содержится:

$$I_T = \{w_i, N_i\}, i = \overline{1, |W_T|}.$$

3. Построение обратного индекса I_{DB} БД текстов: создание словаря W_{DB} канонических слов корпуса текстов, включающего все тексты БД, с указанием для каждого канонического слова $w_j \in W_{DB}$ множеств $N_j^{(1)}, N_j^{(2)}, ..., N_j^{(k_j)}$ всех номеров тех предложений каждого текста $T_j^{(m)}, m = \overline{1, k_j}$, из БД, в которых это слово содержится:

$$I_{DB} = \{w_j; N_j^{(1)}, N_j^{(2)}, ..., N_j^{(k_j)}\}, j = \overline{1, |W_{DB}|}.$$

4. Пересечение обратных индексов I_T и I_{DB} с целью получения списка W слов, с точностью до синонимии, общих для I_T и I_{DB}, с сохранением для каждого $w_s \in W$ его веса p_s, равного количеству предложений из БД, в которые входит данное и синонимичные ему слова:

$$W = \{w_s; N_s; N_s^{(1)}, N_s^{(2)}, ..., N_s^{(k_s)}; p_s\}, \text{ где } p_s = \sum_{m=1}^{k_s} \left| N_s^{(m)} \right|.$$

5. Сортировка списка W в порядке возрастания весов входящих в него слов.

6. Распознавание во входном текстовом документе T предложений, заимствованных из текстовых документов БД.

6.1 Пошаговый выбор из списка W очередного слова w_s и его поиск (фиксирование) в каждом предложении текста T, определяемом по номеру из множества N_s; начисление предложению накапливаемых веса p', равного количеству таких слов в нем, и множества весов, каждый из которых, обозначим его p'', равен количеству всех слов данного предложения, а также им синонимичных слов, входящих в одно и то же предложение БД, определяемое одинаковым значением его номера из множеств $N_s^{(k_s)}$; сохранение только тех весов p'' и соответствующих номеров из множеств $N_s^{(k_s)}$, для которых, начиная с $p' > \mu$, $p' - p'' \le \mu$.

6.2 Как только $p'' = l - \mu$, то данное предложение из T является заимствованным из соответствующего текстового документа БД.

7. Конец.

В представленном алгоритме l – количество слов предложения из T, μ – пороговое значение, т.е. максимально допустимое количество слов предложения из T, не входящих в сравниваемое предложение из БД. Сортировка списка W (шаг 5) и использование весов p' и p'' (шаги 6.1, 6.2) существенно оптимизируют алгоритм решения задачи. Действительно, пошаговая обработка отсортированного списка W позволяет сначала обнаружить в предложениях из T заимствованные из БД слова с низким значением веса p_s, что характерно для слов с высокой предметной смысловой нагрузкой. А используемое при этом условие $p' - p'' \le \mu$, начиная с некоторого момента, настолько сужает, как показали эксперименты, множества $N_s^{(k_s)}$, что последующий анализ слов из списка W с большим значением веса p_s, т.е., как правило, общеупотребительных слов, становится уже нетрудоемким.

Выводы

Представленные выше результаты были успешно реализованы в виде прототипа системы автоматического распознавания воспроизведенных фрагментов текстового документа, разработанного для высшей аттестационной комиссии Республики Беларусь, позволившего обеспечить проведение автоматического анализа в белорусско-русскоязычной информационной среде диссертационных работ и научных статей с целью распознавания в них случаев заимствования результатов других авторов как из заранее заданной полнотекстовой базы данных, так и из полнотекстовой БД, полученной в результате автоматического Интернет-поиска по ключевым словам, выделенным из рецензируемых диссертационных работ. Его функциональность обеспечивается развитым лингвистическим процессором, встроенной системой машинного перевода текстовых документов с белорусского языка на русский и дружественным интерфейсом пользователя-эксперта.

Важно, что предложенные решения позволяют системе, построенной в соответствии с ними, обладать преемственностью (путём наращивания мощности используемой лингвистической базы знаний), т.е., в данном случае, способностью порождения новых её версий как с точки зрения поддержки работы с другими языками, так и увеличения глубины распознавания неявного заимствования за счёт использования уровня семантического анализа языка.

Литература

1. Большой энциклопедический словарь [Электронный ресурс]. – 2012. – Режим доступа : http://mirslovarei.com/bes_a/ – Дата доступа: 22.04.2012.
2. Совпель И.В. Система автоматического извлечения знаний из текста и её приложения / И.В. Совпель // Искусственный интеллект. – 2004. – № 3. – С. 668-677.
3. Совпель И.В. Автоматическое распознавание причинно-следственных отношений в текстовых документах / И.В. Совпель // Искусственный интеллект. – 2005. – № 4. – С. 646-650.
4. Крапивин Ю.Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов / Ю.Б. Крапивин // Вестник БрГТУ : Физика, математика, информатика. – 2009. – № 5 (59): – С. 120-123.
5. Крапивин Ю.Б. Автоматическое определение языка текстового документа для основных европейских языков / Ю.Б. Крапивин // Информатика. – 2011. – № 31 июль-сентябрь. – С. 112-116.
6. Воронков Н.В. Методы, алгоритмы и модели систем автоматического реферирования текстовых документов : дис. … канд. тех. наук. / Воронков Н.В. – Мн., 2007. – 165 с.
7. Google [Электронный ресурс]. – 2012. – Режим доступа : http://www.google.com/ – Дата доступа: 22.04.2012.
8. Robertson S. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF / S. Robertson // Journal of Documentation. – 2004. – № 60 (5). – P. 503-520.

Literatura

1. Bolshioj enciklopedicheskij slovar’. 2012. http://mirslovarei.com/bes_a/
2. Sovpel I.V. Iskusstvennyj intellect. 2004. №3. S. 668-677.
3. Sovpel I.V. Iskusstvennyj intellect. 2005. №4. S. 646-650.
4. Krapivin Y.B. Vestnik BrGTU: Fizika, matematika, informatika. 2009. № 5 (59). S. 120-123.
5. Krapivin Y.B. Informatika. 2011. № 31 ijul’-sentjabr’. S. 112-116.
6. Voronkov N.V. Metody, algoritmy i modeli sistem avtomaticheskogo referirovanija tekstovyh dokumentov: Dis. kand. teh. nauk. Mn. 2007. 165 s.
7. Google [Elektronnyj resurs]. 2012. http://www.google.com
8. Robertson S. Journal of Documentation. 2004. № 60 (5). P. 503-520.

RESUME

Y.B. Krapivin

Automatic Recognition of the Fragments of the Text Documents Adopted from the Internet-Available Resources

In the article, automatic recognition of the fragments of the text documents adopted from the Internet-available resources is considered. Besides, the article presents a definition of plagiarism identification problem, as well as description of the system of the automatic recognition of reproduced fragments of the text documents, which uses the corresponding algorithm of their detection and consists of the next main subsystems: the subsystem of the identification of the language of the text document, the subsystem of the machine translation, the automatic indexing and retrieval of the relevant documents subsystem and the subsystem of the identification of the equivalence of the fragments of the documents.

The subsystem of the automatic indexing and retrieval of the relevant documents provides the possibility of the retrieval of the documents relevant to the input document in the given full-text database as well as Internet search with the key words automatically marked out from the analyzed document using TF-IDF method. The subsystem shows acceptable results sending to the Google search engine fifteen key words queries and downloading first fifty Internet-available documents.

Статья поступила в редакцию 31.05.2012.